T4056_김찬호: 개인 회고

- **나는 내 학습목표를 달성하기 위해 무엇을 어떻게 했는가?**
  - **우리 팀과 나의 학습목표는 무엇이었나?**

  우리 팀의 첫번째 학습목표는 지난 대회를 경험하면서 느꼈던 문제인 Github 활용과 팀 단위의 협업을 개선하는 것이었다. 그 외의 학습목표로는 대회 점수에 너무 매몰되지 않고 성장하는 것을 지향했다.

  - **개인학습측면**

  지난 대회를 경험하면서 가장 중요하면서 스스로 부족한 부분이 모델 내부 구조에 대한 이해와 데이터에 대한 이해(EDA)라고 느꼈다. 모델 내부 구조에 대한 이해를 쌓기위해 모델 구현을 시도했고, 또 내부 구조를 수정하면서 학습 성능의 변화를 관찰했다. 데이터에 대한 이해 역시 주어진 베이스라인과 다른 EDA 예시들을 보면서 자체적으로 가설을 세우고 EDA를 진행하기 위해 노력했다.

  - **공동학습측면**

  우리 팀은 기본적으로 각자 개별적으로 모델을 탐색하고 학습하는 방식으로 대회를 진행했고, 학습한 내용과 각자의 방향성에 대해 피어세션과 데일리스크럼 때 공유하고 노션에서 대회 페이지를 따로 생성해서 기록을 남기고 공유했다. 비슷한 계열의 모델을 학습하는 팀원들을 데이터 엔지니어링 등 다양한 방법들을 공유하면서 진행했다.

- **나는 어떤 방식으로 모델을 개선했는가?**
  - **사용한 지식과 기술**

  먼저 EDA를 통해서 얻은 정보와 문제풀이라는 도메인에 대한 지식을 활용해서 데이터 엔지니어링을 진행했다. 그리고 bert 모델을 주로 다루면서 유저 기준으로 시퀀스 데이터를 형성하면 시퀀스의 개수가 적고, 각 시퀀스의 길이 차이가 많이 나서 max_seq_len에 의해 버려져 사용되지 않는 데이터가 많았다. 그래서 더 작은 단위의 여러 개의 시퀀스로 분할하면서 추가적인 정보를 담기 위해 유저 ID와 테스트 ID를 같이 사용하여 그룹화하여 시퀀스를 생성했다. 위와 같은 방법으로 그룹화하니 시퀀스의 갯수가 많이 증가하고 시퀀스 길이의 차이도 많이 줄어들어 모델의 성능이 향상되었다. 이후 Bert 모델에 FE를 진행하고 LightGCN의 그래프 임베딩을 입력으로 사용해보는 등 여러 가지 접근법

을 활용해봤으나 유의미한 성능 향상은 없었다. 이후 density 분포와 모델의 특성을 참고하여 이질적인 모델들끼리 앙상블을 진행하였다.

- **내가 한 행동의 결과로 어떤 지점을 달성하고, 어떠한 깨달음을 얻었는가?**

위와 같은 과정을 통해서 얻은 bert 모델이 우리 팀의 주력 모델로 최종 결과를 위한 앙상블에 사용되었고, 다른 모델들과 조합하여 더 나은 성능을 보여주었다.

이번 대회를 통해 가장 크게 느낀점은 추천 시스템 문제 해결에서 도메인에 대한 이해와 데이터 엔지니어링이 가장 큰 영향을 끼친다는 것이다. 많은 모델에서 성능 향상에 크게 기여한 feature들은 대부분 도메인에 대한 이해에서 비롯된 것이었고, bert 모델의 경우도 데이터 시퀀스를 수정했을 때 가장 큰 성능 향상폭을 보여줬다. 또 처음으로 시계열 데이터를 다루면서 시계열 데이터를 다룰 때 고려해야할 점과 방법을 알 수 있었다.

- **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

먼저 지난 첫번째 대회에서 깃허브에 충돌이 자주일어나서 협업이 제대로 이루어지지 않은 부분을 문제로 설정하고 이를 해결하기위해 모든 파일을 공유하는 것이 아니라 자신이 작업한 코드 일부만 공유하기로 했다. 이러한 방식으로 깃허브에 충돌이 일어나지는 않았지만, 각자가 실험을 진행하면서 각 과정을 깃허브에 올려 공유하지 않고, 실험이 끝난 결과만 올리게되어 실시간 교류가 잘되지 않았다.

또한, 새로운 팀을 구성하자마자 대회가 진행되었고, 지난 대회에 비해 기간이 길어 각 팀원이 각자 자유롭게 모델을 탐색하고 학습을 진행했는데, 이러한 방식은 여러 모델과 실험을 진행할 수 있다는 확장성에 강점이 있지만, 중간에 팀적으로 정체기가 왔을 때 확실한 팀적인 방향성이 없어 문제해결에 어려움을 겪었다.

또, 데이터 분석(EDA) 기반의 접근방법을 시도해보았지만, 아직 익숙치 않아 다양한 유의미한 정보를 빠뜨리는 등의 문제가 있었다.

이러한 많은 아쉬운 점을 겪었지만, 이를 바탕으로 개선방향을 설정해서 나아가야 할 것이다.

- **마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?**

개인적으로 가장 아쉬운 점은 대회 진행 중간에 모델의 디테일한 부분을 수정하는 것에 빠져서 더 큰 그림에서 방향을 설정하는 것이 힘들었던 것이다. 부스팅 모델에서는 데이터 피쳐를 다양하게 생성해서 넣어주는 것이 성능 향상에 큰 의미가 있고, 시계열 모델에서는 시퀀스 데이터의 양을 늘려주는 것이 성능 향상에 큰 의미가 있

다는 점에 주목해서 방향 설정을 했다면 더 큰 범위에서 유의미한 실험들을 많이 시도하고 더 큰 성능 향상을 이루어낼 수 있었을 것 같다.

ex) 비선형 함수를 통해 데이터 분포 변경, 데이터 어그멘테이션(noise..), correlation 활용, 그룹화를 통해 파생 변수 만들기, 시계열적인 요소를 고려해서 다양한 파생 변수 만들기

- **한계/교훈을 바탕으로 다음 프로젝트에서 스스로 새롭게 시도해볼 것은 무엇일까?**

먼저 low level에서의 모델 이해와 성장을 위해 모델 템플릿을 수작업으로 직접 진행해보는 것이 시간과 노력을 들어가겠지만, 성장에 큰 도움이 될 것 같다. 또, 이번 대회는 4주라는 기간을 활용하기 위해 각자 자유롭게 탐색을 진행했지만, 아쉽게도 한 팀으로써의 방향 설정과 교류는 조금 부족했던 것 같아 다음 프로젝트에는 한 팀으로 진행할 수 있는 시스템을 구축해야할 것 같다.

또, Github를 활용한 협업을 보완하려고 목표를 설정했지만, 이번 대회에서도 제대로 지켜지지 않아 실제 현업이나 다른 팀에서 어떤 식으로 협업하는지 살펴보고 우리 팀 내에 도입해야할 것 같다.

마지막으로 너무 디테일한 문제에 매몰되지 않고, 데이터 위주의 사고방식을 통해 [비선형 함수를 통해 데이터 분포 변경, 데이터 어그멘테이션(noise..), correlation 활용, 그룹화를 통해 파생 변수 만들기, 시계열적인 요소를 고려해서 다양한 파생 변수 만들기] 등의 방법론으로 데이터 기반의 접근방법을 통해 성능향상을 노려봐야할 것 같다.